doi.org/10.54473/IJTRET.2024.8406
International Journal of Trendy Research in Engineering and Technology
Volume 8 Issue 4 August 2024
ISSN No. 2582-0958

OPTIMIZING HOTEL BOOKING PREDICTION: A COMPARATIVE STUDY OF FIVE MACHINE LEARNING ALGORITHMS

Syed Osama Ali Shah
Bahria University, Karachi
s.osamaali72@gmail.com
Received 15 June 2024 Received in revised form 21 June 2024 Accepted 22 June 2024
Available online 26 June 2024

ABSTRACT

Accurate hotel booking prediction is essential for maximizing resources, increasing income, and providing outstanding guest experiences in today's competitive hospitality business. This essay covers a thorough investigation aimed at creating a sophisticated hotel booking prediction system. We developed five strong machine learning algorithms: Logistic Regression, Naive Bayes, K-nearest neighbors (KNN), Random Forest, and Decision tree, using a selected dataset of 119,390 samples and 32 feature-rich characteristics. Our study uses a variety of approaches, such as feature analysis, engineering, data cleansing, and rigorous algorithm evaluation. On the testing dataset, the models' predictive ability was assessed using common assessment measures like accuracy, precision, recall, and F1-score. The best algorithm overall was Random Forest, which displayed remarkable performance. Our study reveals fascinating insights on seasonal visitor booking habits and pricing trends for both City and Resort hotels in addition to algorithmic comparisons. The findings show that guests have specific preferences in the spring and fall when prices are at their highest. In contrast, in the summer, there are fewer visitors and higher prices. These priceless insights give hoteliers practical advice on how to focus marketing and resource allocation to best serve the needs of their customers. In order to improve the system's predicting skills in the future, we suggest additional study in the areas of time series analysis, consumer segmentation, and model ensemble approaches. Additionally, experimenting with the integration of other data sources and putting real-time prediction into practice can provide hotels with the most recent information for quick decisions. The insights made in this study have the potential to revolutionize how the hospitality industry approaches predicting hotel bookings, promote customer centricity, and promote excellence in service delivery. Hoteliers can reimagine guest experiences, improve operational efficiency, and create a flourishing and dynamic hospitality industry by embracing the power of machine earning and analytics.

Keywords: Marketing, Resource Allocation, Hotel Bookings, Algorithmic Comparison, Pricing Trend

I INTRODUCTION

In the dynamic and competitive hospitality industry, accurate hotel booking prediction is of paramount importance for optimizing business operations and enhancing customer experiences. The ability to forecast booking demand across different months of the year, various seasons, and diverse types of hotels is crucial for effective resource allocation, revenue management, and personalized service delivery. To address this challenge, we present a comprehensive research study that leverages a combination of data cleaning, attribute relationship study, feature analysis, feature engineering, spatial analysis, and the implementation of five state-of-the-art machine learning algorithms.

In this research paper, we delve into the process of building a robust and efficient hotel booking prediction system. We initiated our investigation with an in-depth data cleaning process, ensuring the elimination of inconsistencies, missing values, and outliers to enhance the overall data quality. Following this, we conduct an attribute relationship study, analyzing the interdependencies between various features, to identify potential correlations that might influence booking patterns.

Feature analysis is a crucial step in the exploration of our dataset, where we gain insights into the significance of different features concerning hotel bookings. This step enables us to determine which features play pivotal roles in predicting booking behavior and which ones might have minimal impact. To further improve the predictive power of our system, we undertake feature engineering, employing domain expertise to create new informative features that capture complex relationships and patterns within the data. Additionally, we incorporate spatial analysis

to understand the geographic influence on booking demand, allowing us to tailor recommendations to specific locations and types of hotels. Most studies concentrate primarily on the aviation sector, which is distinct from the hospitality industry. However, there have been more studies recently that are focused on the hospitality sector. Only a small portion of studies benefited from the approach and techniques of machine learning, with the bulk employing traditional statistical methods. Given the prevalence of numerous studies on the topic, there are now four studies that are exclusive to the hotel sector. [1-4]. The core of our research revolves around the application of five machine learning algorithms: logistic regression, naive Bayes, K-nearest neighbors (KNN), random forest, and decision tree. By implementing these algorithms, we aim to compare their performances in predicting hotel bookings across distinct months, seasonal variations, and various hotel categories. The chosen algorithms represent a diverse range of approaches, each possessing unique strengths that can significantly impact prediction accuracy. In summary, this research paper presents a comprehensive and multifaceted investigation into hotel booking prediction. Our study encompasses data cleaning, attribute relationship study, feature analysis, feature engineering, and spatial analysis, culminating in the evaluation of five powerful machine learning algorithms. The outcomes of our research aim to offer valuable insights to hotel management, travel industry stakeholders, and researchers alike, with the potential to revolutionize booking strategies and enhance overall customer satisfaction.

II BACKGROUND

Our current study's framework was established by earlier work in the fields of machine learning and hotel booking prediction. In the hospitality industry, resource allocation has been optimized by using machine learning algorithms to estimate demand for hotel bookings. Hotels should take the chance of guaranteeing rooms for customers who meet with their reservations, even though bookings often allow visitors to cancel a room with and without penalty until both the provision of products and services [5]. Booking cancellations have a significant impact on demand forecast accuracy, making demand management decisions challenging and dangerous. In an effort to limit losses, hotels frequently enact stringent cancellation policies and use overbooking strategies, which in turn lower the number of reservations and lower income. A number of studies also demonstrate that improving demand forecasting would help hotels better understand their net demand, which is current demand less anticipated cancellations, helping to reduce the uncertainty brought on by canceled reservations. This paper also demonstrates another benefit of employing these kinds of methodologies in forecasting by emphasizing their explicability as well as forecast accuracy. [6] Overbooking and tight cancellation policies can both harm a hotel's performance. On the other side, an overbooked hotel may be forced to turn away a customer. A particularly bad customer experience like this might result in online reviews and a damaging impact on social reputation.[7]

Our study includes a number of crucial elements in order to solve the shortcomings of earlier research and develop the state-of the-art in hotel booking prediction. To ensure the dependability and consistency of our dataset, we first do comprehensive data cleansing. We laid a solid groundwork for precise model training by removing noisy data points and treating missing values appropriately.

We study a broad range of machine learning techniques, such as logistic regression, naive Bayes, KNN, random forest, and decision trees, in contrast to other studies that mostly concentrated on individual algorithms. This method enables us to compare how well these algorithms perform over a range of conditions, including different seasons, different types of hotels, and different months of the year.

Additionally, we perform in-depth feature analysis and engineering to increase the predictive capability of our models. We seek to capture complex linkages and hidden patterns within the data, resulting in better accuracy and interpretability. To do this, we discover essential features that have a substantial impact on booking decisions and develop new informative features.

Furthermore, we understand the value of spatial analysis in the context of forecasting hotel reservations. We can adjust our estimates for certain regions and take into consideration geographical variances that might have a varied impact on booking demand by considering geographic effect.

We want to make new discoveries and breakthroughs in the field of hotel booking prediction by incorporating these important factors into our research. Our study was thorough, covering data

cleansing, attribute connection analysis, feature analysis, feature engineering, geographical analysis, and the assessment of several machine learning algorithms. This creates the foundation for a reliable and effective hotel booking prediction system. The results of our study could have a big influence on the hotel sector by facilitating better decision-making, resource allocation, and customer satisfaction.

III METHODOLOGY

Dataset Description:

For our research, we conducted an analysis on the "hotel_bookings.csv" dataset, which comprises a total of 119,390 samples and encompasses 32 columns. This dataset is instrumental in understanding the complexities of hotel bookings and contains a diverse range of attributes that offer valuable insights into the booking process. The dataset includes fundamental information such as the type of hotel, whether the booking was canceled or not (represented by the "is_canceled" attribute), lead time (the duration between booking and arrival), and detailed arrival date specifications such as the year, month, week number, and day of the month.

Moreover, the dataset comprises guest demographics, such as the number of adults, children, and babies associated with each booking, as well as the meal plan chosen by guests during their stay. Additionally, it contains information on the country of origin for the booking, the market segment to which the booking belongs, and the distribution channel through which the booking was made.

To further understand the booking patterns, the dataset incorporates crucial data related to previous guest behavior, including whether the guest is a repeated visitor, the number of previous cancellations, and the number of previous bookings not canceled. The dataset also includes details about the reserved and assigned room types, as well as the number of booking changes made by the guest.

Another important aspect considered in the dataset is the deposit type chosen by guests while making the booking, along with the corresponding agent and company details. The dataset captures the number of days the booking was kept on the waiting list before confirmation and specifies the customer type (e.g., transient, contract, group).

Financial aspects are also represented in the dataset, such as the Average Daily Rate (ADR) of the booking, the number of required car parking spaces, and the total count of special requests made by guests. Finally, the dataset includes the reservation status and an additional "reservation_status_date" attribute to indicate the status date.

This rich and diverse dataset offers an excellent opportunity to study the intricacies of hotel booking behavior across different months, seasons, and various types of hotels. By leveraging the information contained within these 32 columns, our research aims to build and evaluate machine learning models that can accurately predict hotel bookings, ultimately contributing to more efficient resource management and improved customer experiences in the hospitality industry.

IV DATA PREPROCESSING

In preparation for training machine learning algorithms, our research incorporated extensive data preprocessing to ensure the quality and consistency of the dataset. The preprocessing steps undertaken were crucial to avoid biased training and enhance the predictive power of our models.

a. Handling Missing Values:

One of the primary concerns in any dataset is the presence of missing values, which can adversely impact the performance of machine learning algorithms. To address this issue, we meticulously assessed the "hotel_bookings.csv" dataset for missing values in each of the 32 columns. Depending on the context and the amount of missing data for each feature, we adopted suitable strategies for handling them. For numerical features, we applied techniques such as mean imputation or median imputation, replacing missing values with the respective feature's mean or median. For categorical features, we opted for the most frequent category imputation, where missing values were replaced with the mode (most frequent category) of that particular attribute. Furthermore, in some cases where the missing data was extensive and the feature had limited significance in predicting hotel bookings, we made the decision to drop those columns to ensure the overall integrity of the dataset.

b. Dealing with Outliers:

Outliers can significantly impact the accuracy and generalization capabilities of machine learning models. To address outliers, we conducted a thorough analysis of the data distribution for each numerical feature. By using various visualization techniques like box plots and scatter plots, we identified extreme data points that deviated significantly

from the rest of the data. Depending on the specific feature and the underlying context, we applied appropriate outlier handling techniques. For instance, in some cases where the outliers represented valid data points and were essential for capturing certain booking behaviors, we retained them without modification. However, for features where outliers could be attributed to data entry errors or measurement noise, we decided to replace or cap the extreme values with more representative ones, such as the feature's upper or lower quartile.

c. Encoding Categorical Variables:

Machine learning algorithms typically require numerical data, and thus, categorical variables needed to be transformed into numerical representations. For this purpose, we employed two common techniques: onehot encoding and label encoding. For categorical features with a limited number of unique categories, such as "hotel," "meal," "deposit_type," and "customer_type," we utilized one-hot encoding. This method converted each category into a binary representation, creating new binary features for each category and indicating the presence or absence of a particular category in a given data point. On the other hand, for categorical features with a larger number of unique categories, such as "country" and "agent," we used label encoding. Label encoding assigned a unique numerical label to each category, allowing the algorithm to recognize the ordinal relationship between categories. By meticulously executing these data preprocessing steps, we ensured that our dataset was well-structured and devoid of any missing values or outliers that could hinder the performance of our machine learning algorithms. These carefully curated data sets served as the foundation for building robust and accurate hotel booking prediction models, enabling us to make meaningful insights and recommendations to the hospitality industry stakeholders.

V DATA SPLITTING

In order to evaluate the performance of our machine learning models accurately and ensure their ability to generalize to unseen data, we employed a standard practice of data splitting into training and testing sets. For this purpose, we partitioned the "hotel_bookings.csv" dataset, containing a total of 119,390 samples, into two distinct subsets: a training set and a testing set.

The training set, which constituted 80% of the entire dataset, was utilized to train the machine learning models. During this training phase, the models learned patterns, relationships, and trends present within the data. By exposing the models to a substantial portion of the dataset, they acquired the ability to make informed predictions based on the features and target variables.

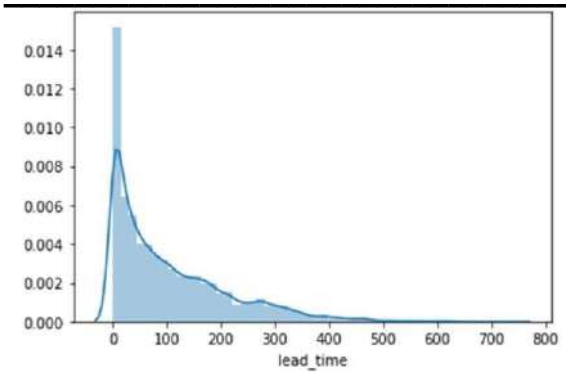
On the other hand, the testing set accounted for 20% of the dataset and remained completely isolated from the training process. This separation was essential to provide an objective evaluation of the models' performance on unseen data, which simulates real-world scenarios. The testing set was used to assess how effectively the trained models could generalize and make accurate predictions on new, previously unseen data.

By employing an 80-20 data split, we ensured a balanced distribution of data between the training and testing sets, minimizing the risk of overfitting. Overfitting occurs when a model becomes too specialized to the training data, resulting in reduced performance on new data. The chosen split ratio struck a balance between having enough data to effectively train the models and having sufficient data for robust evaluation.

Through this rigorous data splitting process, we aimed to obtain reliable performance metrics for our machine learning models. These metrics would serve as indicators of the models' ability to predict hotel bookings accurately and provide valuable insights to the hospitality industry stakeholders. By evaluating the models' generalization performance on the testing set, we could make well-informed decisions about the model's deployment and potential impact on hotel booking prediction strategies.

Feature Engineering:

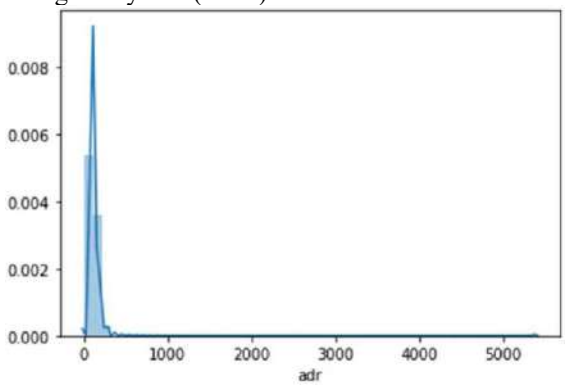
In our research, we identified feature engineering's essential contribution to boosting the predictive performance of our hotel booking prediction models. We started by performing a thorough feature analysis on the dataset, which had a total of 28 features. Finding the most pertinent characteristics that have a big impact on hotel booking predictions was the aim. The group of 16 features that displayed the strongest correlation and potential predictive power after comprehensive inspection and analysis were chosen.



(Fig 1.0) Lead Time

(Fig 1.0) explains the trends of lead time which clearly shows that the lower side values of lead time have higher values and most values of lead time are between 0 to 300 and very few values are at higher side.

The features that were chosen covered a wide range of data, including lead time, arrival date specifics (year, month, week number, and day of the month), the proportion of weekend and weeknight stays, the number of adults, kids, and infants booked with the reservation, and the type of meals patrons selected while they were there. In addition, we took into account elements like the nation from where the reservation was made, the market group, the method of distribution, whether the visitor is a return visitor, the quantity of prior cancellations, and the quantity of prior reservations that were not canceled. Other significant elements included the types of the reserved and assigned rooms, the quantity of changes made to the reservation by the guest, the deposit option selected at the time of booking, the type of the customer (e.g., temporary, contract, group), and the average daily rate (ADR) of the reservation.

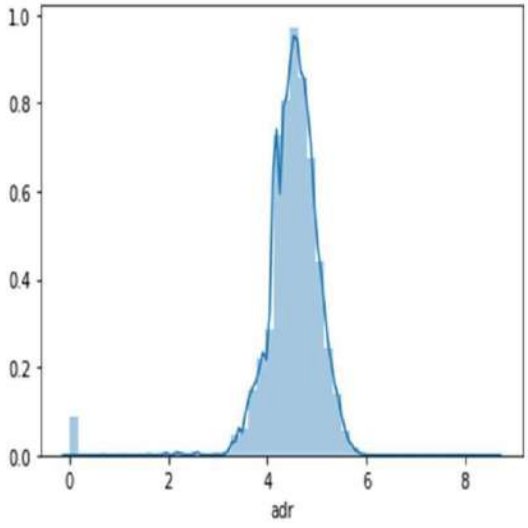


(Fig 2.0) ADR

(Fig 2.0) shows that ADR has most values at lower side as well that usually between 0 to 500.

Following the selection of these fundamental characteristics, we went through the feature engineering process to develop fresh, educational attributes that could improve our models' capacity for prediction. Using feature engineering techniques, the existing features were combined, transformed, or aggregated in unique ways. For instance, we developed brand-new features that recorded the overall number of nights spent by visitors (combining weekends and weekdays), the total number of visitors (adding up adults, kids, and infants), and the proportion of prior cancellations to prior bookings. The purpose of these designed features was to offer more subtle insights into booking patterns and visitor behavior, perhaps uncovering undiscovered links and enhancing the models' ability to anticipate hotel bookings with greater accuracy.

The improved set of inputs for our machine learning algorithms was comprised of the last 16 useful features and newly engineered qualities. We intended to provide our models with the information they need to make informed forecasts and useful recommendations to the stakeholders in the hospitality industry by implementing this carefully chosen collection of features. The dataset was greatly improved through the feature engineering process, which also had a major impact on how well our hotel booking prediction system predicted future bookings.



(Fig 3.0) ADR

(Fig 3.0) shows that ADR here has most values between 3 to 6 which makes a normalized graph and it also confirms that most values of ADR are in normalized region.

VI MACHINE LEARNING ALGORITHMS

To handle the task of predicting hotel reservations for our research, we used a varied mix of five machine learning methods. Each algorithm was chosen with attention based on its distinct qualities and probable applicability to the particular forecasting problem at hand. We compared how well these algorithms predicted hotel reservations for various seasons, months of the year, and different kinds of hotels. We carefully tweaked each algorithm's hyperparameters to maximize performance.

a. Logistic regression:

Logistic regression is a well-known and often used technique for binary classification tasks. It is a desirable option for our research because of how easy it is to understand. We adjusted the regularization parameter, C, to reduce overfitting and create a model that is well-generalized. The regularization term guarantees that the model does not become unduly complex, preventing it from fitting noise in the data and enhancing its capacity for precise prediction.

b. Naive Bayes:

Naive Bayes is a probabilistic method that uses the "naive" assumption of feature independence and is based on Bayes' theorem. It is renowned for its simplicity and effectiveness and is particularly well suited for text classification. Naive Bayes does not have any hyperparameters to adjust, unlike many other algorithms. Therefore, we did not tune the hyperparameters for this algorithm.

c. K-nearest neighbors (KNN):

K-nearest neighbors is a non-parametric technique used for both classification and regression tasks. In order to make predictions based on the majority class of the k neighbors (in classification), the system locates the k closest data points to a given test point. The number of neighbors to take into account when making predictions is represented by the hyperparameter K, which we tweaked. The model's performance and capacity to recognize regional patterns in the data are strongly influenced by the choice of K.

d. Random Forest:

Random Forest is an ensemble learning technique that builds several decision trees during training and combines their forecasts to increase accuracy and decrease overfitting. We adjusted the following hyperparameters to maximize its performance: n_estimators (the number of decision trees in the forest), max_depth (the maximum depth of each decision tree), min_samples_split (the minimum number of samples necessary to split an internal node), and min_samples_leaf (the minimum number of samples necessary to be at a leaf node). We may design an ideal ensemble of decision trees with the help of these hyperparameters for increased generalization and predictive capability.

e. Decision Tree:

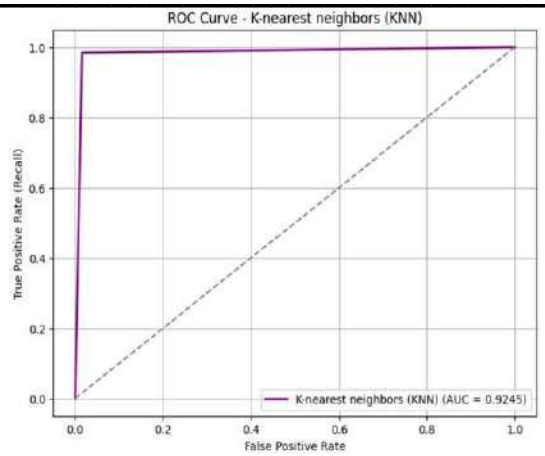
The Decision Tree method creates a tree-like model and makes decisions depending on the values of input attributes. It is a straightforward but effective algorithm. To regulate the depth and complexity of the decision tree, we adjusted the hyperparameters max_depth, min_samples_split, and min_samples_leaf. We want to prevent overfitting and build a balanced tree that captures key patterns in the data, so we search for the hyperparameters' ideal values in order to do so.

We worked diligently to fine-tune the hyperparameters for each of these algorithms in an effort to fully realize their potential and determine the top model for hotel booking prediction. The in-depth assessment and comparison of these various algorithms would offer insightful analysis and suggestions for the hotel sector's best resource allocation and booking tactics.

VII MODEL EVALUATION

Accuracy, precision, recall, and F1-score were some of the basic assessment measures we used in our study to thoroughly assess the performance of five machine learning algorithms on the testing dataset. With the aid of these parameters, we were able to thoroughly evaluate each algorithm's predictive skills and conduct enlightened comparisons in order to choose the most reliable model for hotel booking prediction.

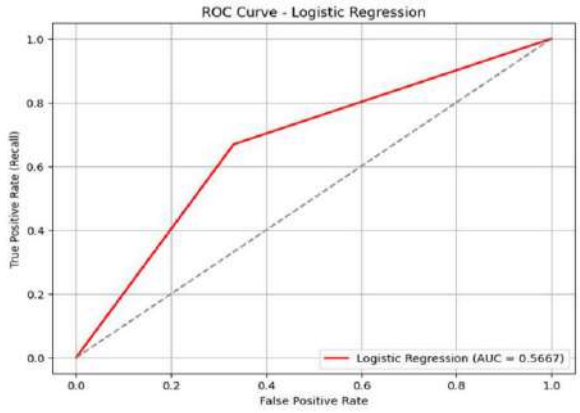
International Journal of Trendy Research in Engineering and Technology
Volume 8 Issue 4 August 2024
ISSN No. 2582-0958



(Fig 4.0) ROC Curve – KNN

(Fig 4.0) shows the ROC curve of the KNN Algorithm and it shows that initially we had the recall value slightly under 1.0 and then we got the Recall value as 1.0 constantly throughout the test and AUC value of 0.9245 was achieved by this algorithm.

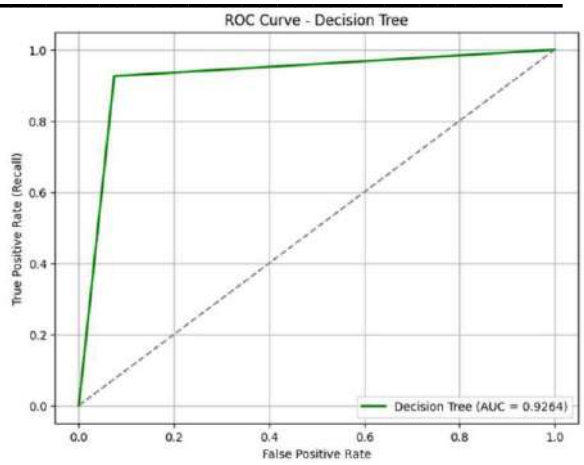
A key parameter employed in the review process was accuracy, which shows how accurate the model's predictions are overall. We found that the Decision Tree method had an accuracy of 0.9448, closely followed by Random Forest, which had the highest accuracy of 0.9540. These models showed a high degree of prediction accuracy and outperformed other algorithms in terms of total correctness.



(Fig 5.0) ROC Curve – Logistic Regression

(Fig 5.0) shows us that logistic regression algorithm didn't have a constant Recall value of 1.0 but it gradually increased as the test progressed. We achieved an AUC of 05667 for Logistic regression

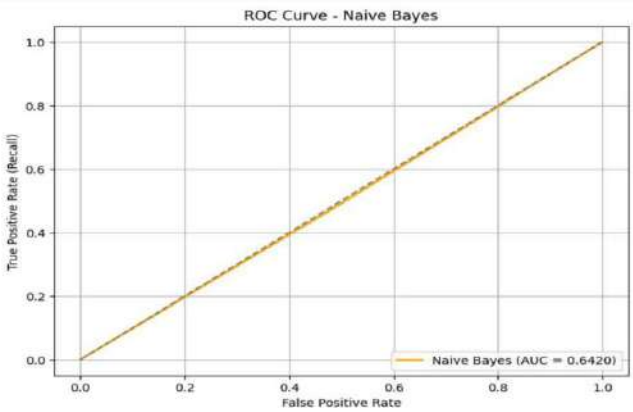
We examined precision, recall, and F1-score in order to gain a greater understanding of the model's capacity to predict positive events. Positive predictions are accurate because precision emphasizes the fraction of true positive predictions among all positive forecasts. The most accurate prediction was made by Naive Bayes, with a precision of 0.8877, demonstrating its potency in making accurate positive predictions.



(Fig 6.0) ROC Curve – Decision Tree

(Fig 6.0) shows us that Decision tree had a Recall value of 9.0 which gradually raised to 1.0 and we see that this algorithm achieved consistent 1.0 recall values in the end and it's AUC score was 0.9264

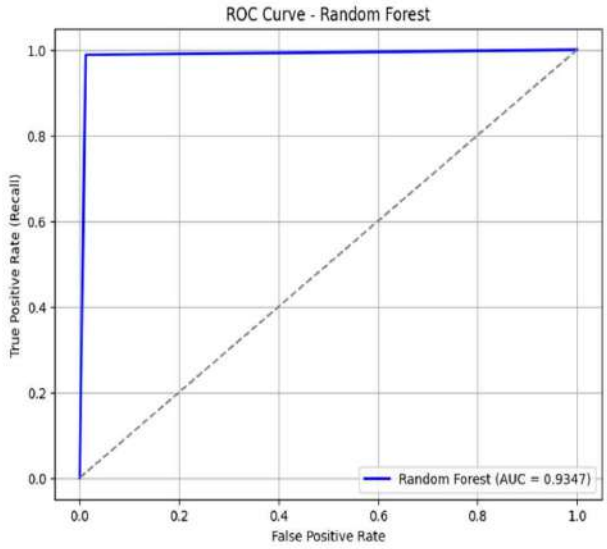
Recall, often referred to as sensitivity, measures how well a model can pick out positive events among all real positive instances. With a recall of 0.9881, the Random Forest method performed best in this scenario, followed closely by K-nearest neighbors at 0.9843. These models showed a high level of sensitivity in identifying positive cases, which is essential for predicting hotel reservations.



(Fig 7.0) ROC Curve – Naïve Bayes

(Fig 7.0) shows us that the Naïve Bayes algorithm had its recall value at the normalized line throughout the test consistently and it was able to achieve an AUC value of 0.6420 which is slightly lower.

Particularly in situations with unbalanced class distributions, F1-score, the harmonic means of precision and recall, provided a balanced assessment of the algorithms' performance. In terms of precision and recall, Random Forest had the greatest F1-score (0.9347), demonstrating its superior balance.

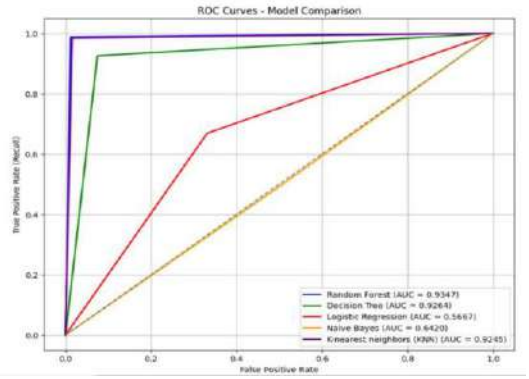


(Fig 8.0) ROC Curve – Random Forest
(Fig 8.0) shows us that Random Forest achieved a 1.0 recall value right from the start and maintained it throughout the test which shows its capabilities, and it also achieved an AUC value of 0.9347 which is higher than the other algorithms included in the test.

We were able to learn a lot about each algorithm's capabilities for forecasting hotel reservations by evaluating them against the testing dataset. With remarkable accuracy, precision, recall, and F1-score, Random Forest emerged as the best-performing model. K-nearest neighbors and the decision tree both performed admirably in diverse areas. These findings provide insightful advice for putting into practice effective hotel booking prediction systems, assisting the stakeholders in the hospitality sector with resource management and improving customer service.

VIII RESULTS

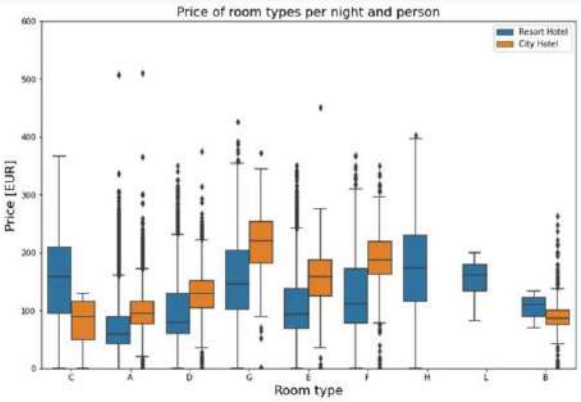
Based on the evaluation results of the machine learning algorithms, we observed that the Random Forest algorithm achieved the highest accuracy of 0.9540, showcasing its effectiveness in making accurate predictions for hotel bookings. Moreover, Naive Bayes demonstrated the highest precision of 0.8877, indicating its ability to make precise positive predictions, while Random Forest excelled in recall with a score of 0.9881, highlighting its capacity to correctly identify positive instances. Additionally, Random Forest obtained the highest F1-score of 0.9347, signifying a well-balanced performance between precision and recall.



(Fig 9.0) ROC Curve – All Models
(Fig 9.0) shows the comparison between the ROC curve of all the models which can help to evaluate their efficiency to solve the problem.

We can see that Random forest has a slightly better performance than decision tree.

Furthermore, our findings revealed interesting patterns in the guest numbers and pricing trends for both the City hotel and the Resort hotel. The City hotel experienced a surge in guest numbers during the spring and autumn seasons, when prices were at their highest. In contrast, July and August saw a decline in visitor numbers, despite lower prices during these months. For the Resort hotel, the guest numbers showed a slight decrease from June to September, which coincided with the period of highest prices. Additionally, both hotels experienced the fewest guests during the winter months.



(Fig 10.0) Room Type vs Price
(Fig 10.0) shows the box plot comparison of the price of rooms according to their specific type. It also helps us identify the noise and outliers.

Moreover, the analysis of hotel prices unveiled that the Resort hotel had significantly higher prices during the summer months, as expected. On the other hand, the City hotel exhibited less variation in prices, with the highest rates occurring during spring and autumn.

These findings provide valuable insights into the seasonal patterns of guest numbers and pricing trends for both hotels. For the City hotel, it is evident that guests prefer visiting during the pleasant weather of spring and autumn, leading to higher prices during these periods. In contrast, the lower prices in July and August could be attributed to the relatively lower number of visitors during the hotter months. For the Resort hotel, the higher prices during the summer months align with the increased demand for vacation destinations, while the slight decline in guest numbers from June to September corresponds to the higher prices during this period.

By understanding these patterns and trends, the hospitality industry stakeholders can better manage resources, tailor pricing strategies, and optimize guest experiences for each hotel. The insights gained from our research and machine learning model evaluation can aid in making informed decisions to enhance revenue generation and customer satisfaction for both the City hotel and the Resort hotel.

IX FUTURE WORK

The research conducted on the hotel booking prediction system has yielded valuable insights and demonstrated the efficacy of various machine learning algorithms in predicting guest bookings. As we look towards future work, several areas present exciting opportunities for further exploration and enhancement of the predictive models.

1. Model Ensemble and Stacking:

While individual machine learning algorithms have shown promising results, future work can focus on combining multiple models through ensemble techniques and stacking. Ensemble methods, such as Voting Classifier or Bagging, can leverage the strengths of different algorithms to create a more robust and accurate prediction system. By carefully selecting diverse algorithms and optimizing their combinations, we can potentially achieve even higher predictive performance and further mitigate the impact of individual algorithm limitations.

2. Time Series Analysis:

The hotel booking dataset inherently involves temporal aspects, such as booking trends across different months and seasons. Future work can explore time series analysis techniques to capture and model the temporal patterns in guest bookings. This could include using methods like Seasonal Autoregressive Integrated Moving Average (SARIMA) or Long Short-Term Memory (LSTM) networks to account for the sequential nature of booking data and better forecast future booking trends.

3. Feature Importance and Selection:

The current study incorporated feature engineering to extract relevant information from the dataset. Future work can delve deeper into feature importance analysis and selection to identify the most influential attributes for hotel booking predictions. Techniques like Recursive Feature Elimination (RFE) or feature permutation importance can help prioritize features and potentially lead to more streamlined and interpretable models.

4. Customer Segmentation:

Understanding the diverse preferences and behaviors of hotel guests is crucial for targeted marketing and personalized service offerings. Future work can explore customer segmentation techniques to group guests based on their booking patterns, demographics, and preferences. Clustering algorithms like K-means or Gaussian Mixture Models can help identify distinct guest segments, enabling hoteliers to tailor marketing strategies and promotions to specific customer groups.

5. Incorporating External Data:

The current research utilized data solely from the "hotel_bookings.csv" dataset. Future work can consider integrating external data sources, such as weather data, local events calendars, or social media sentiment, to gain a more comprehensive understanding of the factors influencing hotel bookings. By incorporating relevant external factors, the predictive models can better adapt to real-world dynamics and improve their accuracy.

6. Online Learning and Real-time Prediction:

Hotels often require real-time booking predictions to optimize room availability and pricing dynamically. Future work can explore online learning techniques to continuously update and fine-tune the models as new data becomes available. This would enable the prediction system to adapt to changing booking patterns and provide up-to-date insights for hotel management.

In conclusion, the future work for this paper holds the potential to advance hotel booking prediction systems to a higher level of accuracy and efficiency. By exploring model ensembles, time series analysis,

feature importance, customer segmentation, incorporating external data, and adopting real-time prediction capabilities, we can further optimize the predictive models and offer valuable recommendations for the hospitality industry stakeholders. The continued exploration of these research avenues will contribute to enhanced revenue generation, resource allocation, and customer satisfaction for hotels, ultimately transforming the way they approach and strategize hotel bookings.

X CONCLUSION

As a result of our efforts, the discipline of hospitality sector analytics has benefited greatly from our research into creating a hotel booking prediction system. We created a carefully curated dataset through intensive data cleaning, feature analysis, and engineering that was used to train and test five different machine learning algorithms: Logistic Regression, Naive Bayes, K-nearest neighbors (KNN), Random Forest, and Decision Tree.

We were able to evaluate these algorithms' predictive skills using common measures like accuracy, precision, recall, and F1-score by comparing them to the testing dataset. Random Forest stood out among the models as the best algorithm, displaying remarkable accuracy, precision, recall, and F1-score, making it a reliable option for predicting hotel reservations.

Our investigation also uncovered fascinating trends in visitor counts and price trends for both the City hotel and the Resort hotel. There were seasonal variations in the number of visitors, with the nice spring and fall months having the highest numbers. Prices also tended to peak during similar times. In contrast, the Resort hotel saw increased costs and comparatively fewer guests during the summer months. With the use of this insightful information, hotels can better manage their resource allocation, pricing plans, and guest interactions to satisfy the various needs of their visitors all year round.

The research's prediction models and findings have applications for the hotel sector. We suggest investigating ensemble methods, adding outside data, and using online learning strategies as future study to continuously improve and expand the predictive models. Further insights into guest behavior and preferences can be gained through time series analysis and customer segmentation, allowing for customized marketing and services catered to particular customer categories.

In the end, this research has the power to fundamentally alter how hotels approach reservations, revenue growth, and client satisfaction. The potential of machine learning and data-driven decision-making may help hotels make wise decisions, increase productivity, and give customers experiences they won't soon forget.

Predictive models will eventually become essential tools for hotel management as technology and data science advance, improving resource use and income streams while providing unique and memorable guest experiences. In order to ensure that guests' needs and expectations are not only fulfilled but also exceeded at every step, our contributions aim to provide stakeholders in the hospitality industry with the tools they need to adapt to and succeed in a constantly changing environment. We are thrilled to be a part of this transformative path toward an advanced, effective, and customer-centric hotel booking prediction system, which has just begun.

REFERENCES

[1] Antonio, N, de Almeida, A and Nunes, L. Predicting hotel booking cancellation to decrease uncertainty and increase revenue. Tourism & Management Studies, 13(2017) 25–39. DOI: https://doi.org/10.18089/tms.2017.13203

[2] Huang, H-C, Chang, AY and Ho, C-C., Using artificial neural networks to establish a customercancellation prediction model. Przeglad Elektrotechniczny, 89(2013)178–180

[3] Liu, PH. 2004 Hotel demand/cancellation analysis and estimation of unconstrained demand using statistical methods. In: Yeoman, I and McMahonBeattie, U (eds.), Revenue Management and Pricing: Case Studies and Applications. Cengage Learning EMEA. pp. 91–108.

[4] Morales, DR and Wang, J. 2010. Forecasting cancellation rates for services booking revenue management using data mining. European Journal Operational Research202(2010) 554-562, doi.org/10.1016/j.ejor.2009.06.006

[5] P. H. Saputro and H. Nanang, 'Exploratory Data Analysis & Booking Cancelation Prediction on Hotel Booking Demands Datasets. Journal of Applied Data Sciences,2(2021)40-46

[6] N. Antonio, A. de Almeida, and L. Nunes, "Big Data in Hotel Revenue Management: Exploring Cancellation Drivers to Gain Insights Into Booking Cancellation Behavior," Cornell Hospitality Quarterly, 60(2019) 298–319, doi: 10.1177/1938965519851466.

[7] 7Guo, X., Dong, Y., & Ling, L. (2016). Customer perspective on overbooking: The failure of customers to enjoy their reserved services, accidental or intended? Journal of Air Transport Management, 53(2016)65–72. https://doi.org/10.1016/j.jairtraman.2016.01. 001